Advanced Regression: Linear and generalised linear models II

Garyfallos Konstantinoudis

Feb 20, 2024

Generalised linear model

- Basic definition
- Technical details on exponential families and GLMs
- Logistic regression and binary outcomes
- Generalised linear models in R

Generalised linear model (GLM)

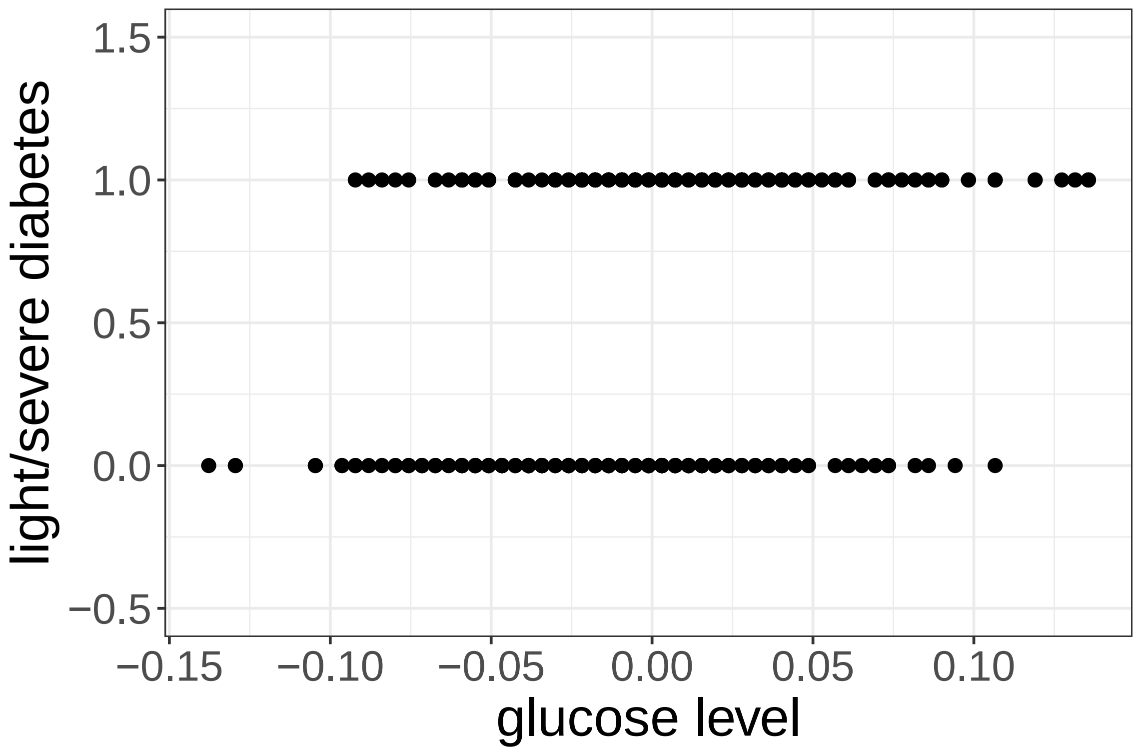
- Linear models can only model a quantitative outcome.
- Quantitative outcomes are defined as a real number, taking possible values from $-\inf$ to $+\inf$.
- Many important data types can by definition not be modelled using a linear model:
  - Dichotomous or binary → only takes two values, 0 or 1
  - Counts → only positive integers (0,1,2,3,...)

> Note
>
> Flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. Residuals are an important quantity for model diagnostics.

Binary outcome and logistic regression

Example: Case-control study

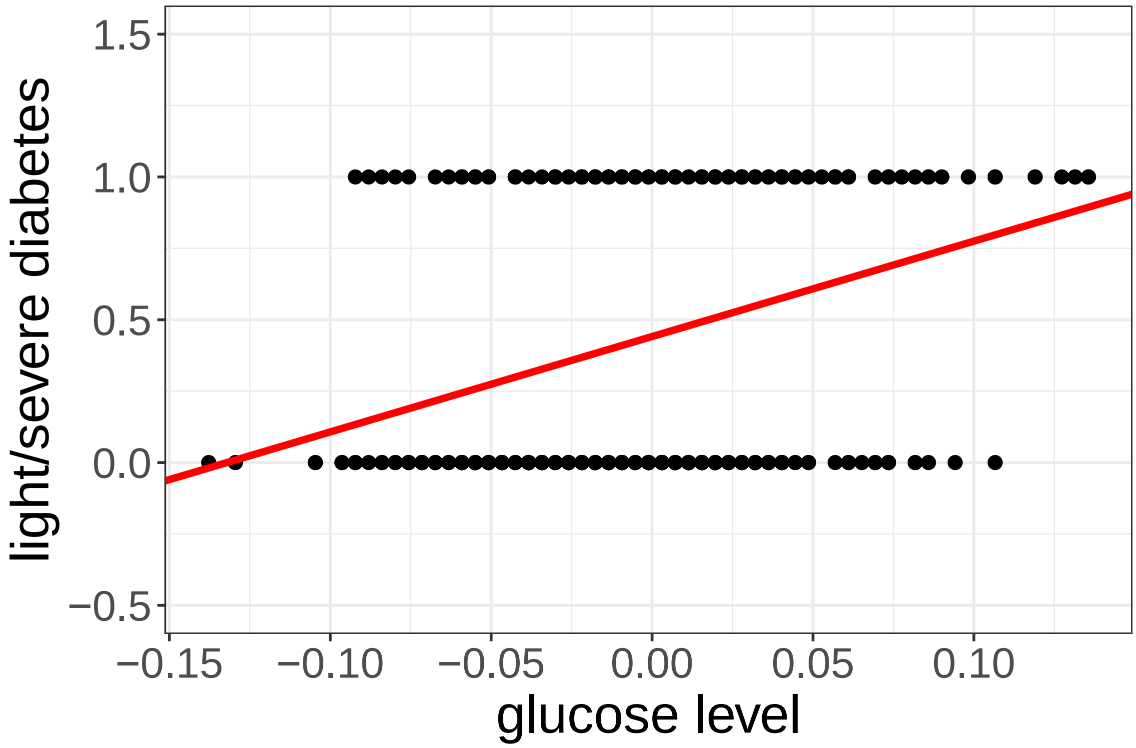
$$y_i = \begin{cases} 1, & \text{If subject } i \text{ is a case} \\ 0, & \text{If subject } i \text{ is a control} \end{cases}$$



Binary outcome and logistic regression

$$y = \underbrace{\alpha + \beta x}_{\text{Linear predictor}} + \epsilon$$

- Linear predictor: $\eta = \alpha + \beta x$ is defined from $-\inf$ to $+\inf$.
- But y only 0 or 1 $\rightarrow$ The linear regression do not match the data well.



How should we model this data?

Key idea 1: Instead of modelling the outcomes ($y = 0$ or $y = 1$) directly, logistic regression models the probability for $y = 1$ denotes as

- $P(y = 1 \mid x)$

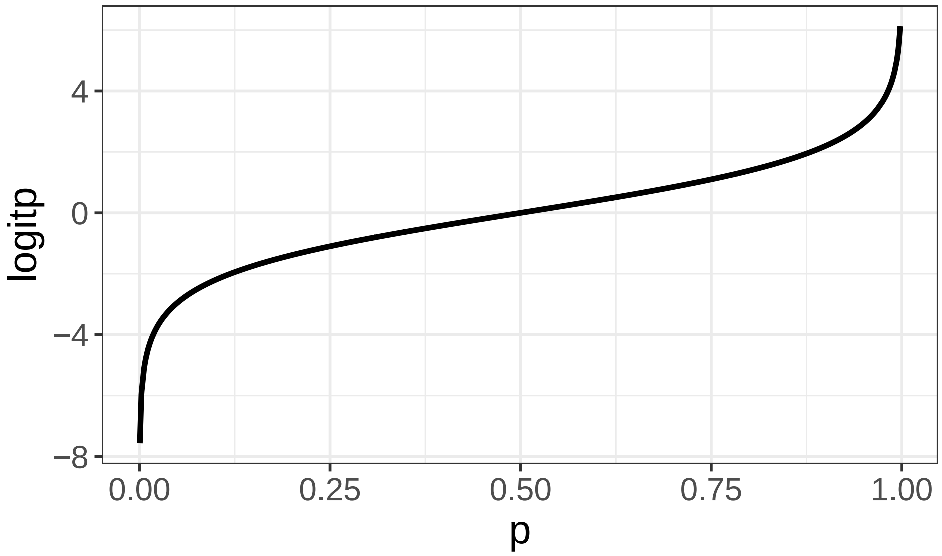
Notes on probabilities for binary data:

- Probabilities can take values from 0 to 1
- Probabilities are symmetric: $P(y = 1 \mid x) = 1 - P(y = 0 \mid x)$

How should we model this data?

Key idea 2: Transform the linear predictor $\eta = \alpha + \beta x_i$ (quantitative, can take values from $-\inf$ to $-inf$) to lie in the Interval $[0, 1]$, which is valid for probabilities.

This can be achieved using the logit function: $\mathrm{logit}(p) = \log(p/(1 - p))$



Logistic regression

$$\text{logit}(P(y=1 \mid x)) = \log(P(y=1 \mid x))/(1-P(y=1) \mid x) = \alpha + \beta x$$

- **Interpretation**: The regression coefficient β in logistic regression represents the **log odds ratio** between $y=0$ and $y=1$.
- **Estimation**: Maximum likelihood

Technical details

- Many important outcome types can be accommodated by GLMs.
- Each of these distributions has a location parameter, e.g. μ for the Gaussian, p for the Bernoulli and Binomial.
- The natural link function between the location parameter and the linear predictor can be derived from the mathematical form of the distribution.

Response	Distribution	E(y)	Link (g)
Continuous	Gaussian	μ	1 (identity)
Dichotomous	Bernoulli	p	logit
Counts	Poison	λ	log

https://en.wikipedia.org/wiki/Generalized_linear_model

Technical details: GLM

The GLM consists of three elements:

1. A probability distribution from the exponential family. Note: Only distributions that can be formulated as an exponential family can be modelled as GLM.
2. A linear predictor $\eta = x\beta$
3. A link function g such that $E(y) = \mu = g^{-1}(\eta)$

Technical details: Exponential families

An exponential family is a set of probability distributions of the following form:

$$f_x(x \mid \theta) = h(x) \exp\{\eta(\theta) \times T(x) - A(\theta)\}$$

where

- θ is the parameter of interest.
- $T(x)$ is a sufficient statistic.
- $\eta(\theta)$ is the natural parameter or link function.

Gaussian distribution as exponential distribution

Gaussian distribution with unknown μ, but known σ:

$$f_\sigma(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\}$$

- $\theta = \mu$
- $h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\}$
- $T(x) = \frac{x}{\sigma}$
- $\eta(\mu) = \frac{\mu}{\sigma}$
- $A(\mu) = \frac{\mu^2}{2\sigma^2}$

Logistic regression and binary outcomes

Binomial distribution with known number of trials n, but unknown probability p:

$$\begin{aligned} f(x \mid p) &= \binom{n}{x} p^x (1-p)^{n-x} = && (1)\\ &= \binom{n}{x} \exp\{x \log(\frac{p}{1-p}) + n \log(1-p)\} && (2) \end{aligned}$$

- $\theta = p$
- $h(x) = \binom{n}{x}$
- $T(x) = x$
- $\eta(p) = \log(\frac{p}{1-p})$
- $A(p) = -n \log(1-p)$

Logistic regression and binary outcomes

Formulate model: Three elements

1. Error distribution for response variable
2. Linear predictor
3. Link function

The three elements of the logistic regression model are:

1. The Bernoulli probability distribution modelling the data: $P(y_i = 1 \mid x_i) = p_i$
2. The linear predictor: $\alpha + \sum_{j=1}^{p} \beta_j x_{ij}$
3. The link function g associating the mean of y, $P(y_i = 1 \mid x_i)$ to the linear predictor: here the link is the logistic link as we set $g(P(y_i = 1 \mid x_i)) = \text{logit}(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

glm() in R

- GLMs can be called in R just as linear models.

```r
library(lars)
library(dplyr)

data(diabetes)
x <- as.data.frame.matrix(diabetes$x)
y <- ifelse(diabetes$y > mean(diabetes$y), 1, 0)
```

```r
glm(y ~ age + sex + bmi + map + ltg, data = x, family = binomial) %>% summary()
```

```
Call:
glm(formula = y ~ age + sex + bmi + map + ltg, family = binomial,
    data = x)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3164     0.1190  -2.659  0.00783 **
age          -1.0206     2.7309  -0.374  0.70860
sex          -5.4254     2.6315  -2.062  0.03923 *
bmi          14.5079     3.0223   4.800 1.58e-06 ***
map          11.8803     2.9652   4.007 6.16e-05 ***
ltg          18.6940     3.1954   5.850 4.91e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 606.61  on 441  degrees of freedom
Residual deviance: 434.17  on 436  degrees of freedom
AIC: 446.17

Number of Fisher Scoring iterations: 4
```

glm() in R

- Different types of exponential families can be called using the family option:
  - binomial(link = 'logit')
  - gaussian(link = 'identity')
  - Gamma(link = 'inverse')
  - poisson(link='log')
- There are similar return values as for the lm function:
  - coefficients
  - residuals

- fitted.values
- linear.predictors: the linear fit on the link scale

Making predictions

1. Train the prediction rule
2. Derive predictions on the linear predictor scale for the new data

```
library(lars)
library(dplyr)
library(ggplot2)
library(patchwork)
```

```
Warning: package 'patchwork' was built under R version 4.3.2
```

```
set.seed(11)

data(diabetes)
x <- as.data.frame.matrix(diabetes$x)
y <- ifelse(diabetes$y > mean(diabetes$y), 1, 0)

glm_predict <- glm(y ~ glu, data = x, family = binomial)
xnew <- data.frame(glu = rnorm(n = 1000, mean = 0, sd = 0.5))
xnew %>% head()
```

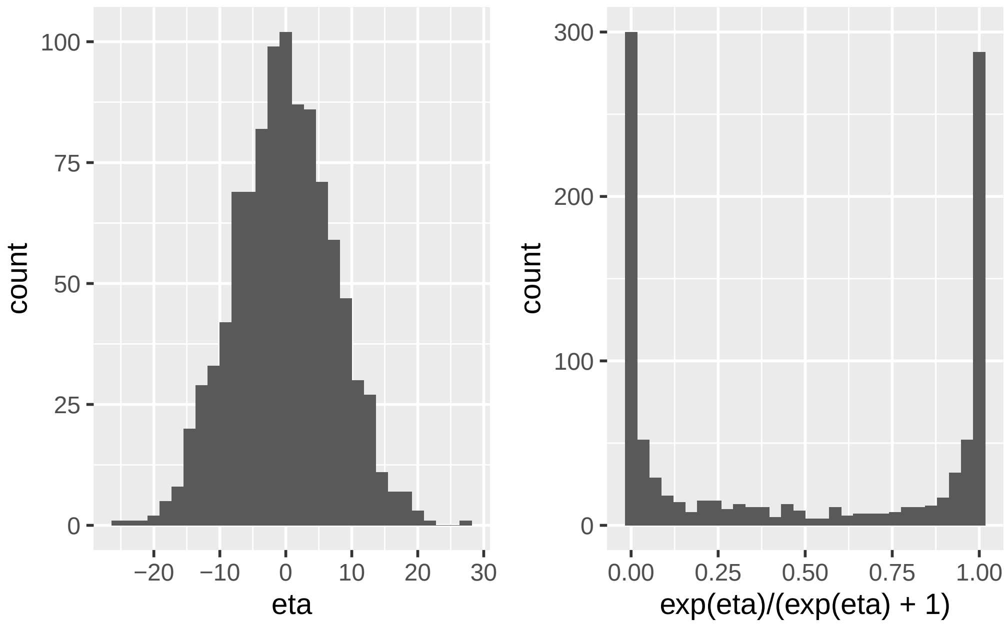
```
          glu
1 -0.29551555
2  0.01329718
3 -0.75827655
4 -0.68132667
5  0.58924458
6 -0.46707566
```

```
eta <- predict.glm(glm_predict, xnew)
```

Plot the predictions

```
ggplot() +
  geom_histogram(aes(x = eta)) |
  ggplot() +
    geom_histogram(aes(x = exp(eta) / (exp(eta) + 1)))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
rbinom(
  n = length(eta),
  size = rep(1, length(eta)),
  prob = exp(eta) / (exp(eta) + 1)
) %>% head()
```

```
[1] 0 0 0 0 1 0
```

Take away: Generalised linear model

The model formulation in GLMs consists of three elements:

1. Error distribution for response variable
2. Linear predictor
3. Link function

Most common data types can be modelled using GLMs

- Continuous $\rightarrow$ Gaussian distribution
- Dichotomous or binary $\rightarrow$ Bernoulli distribution
- Counts $\rightarrow$ Poisson or Binomial (with known number of trial) distribution